Autonomous Vocal and Backing Track Mixing (Proposal)

kelian Li
Center for Music Technology
Georgia Institute of Technology
Atlanta, USA
kli421@gatech.edu

I. Research Statement / Problem

Music mixing is the process of blending multitrack recordings by manipulating the musical sources through audio effects and achieves better audio quality. The quality is ultimately judged by the preference of listeners. Automatic mixing aims to reach a similar level of quality by computational systems without human interventions [1].

Automatic mixing is a fairly new research field that can possibly, or has already, changed the music production process for engineers and amateur music creators. The research can not only increase the efficiency, but also help understand mixing from the aspects from science and arts which may lead to innovations in music production.

This research proposal intends to solve the problem of vocal and backing track mixing through the data-driven method. Previous research either attempted to create single unified general-purposed systems which has not yet succeed, or targeted on a specific scenario like audio mastering. The proposed project focus on vocal and backing track mixing which has not be addressed in previous research but with applications in professional music production and public music participation.

II. Motivation

A. Making Beats: A New Market

Derived from hip-hop culture, beats are essentially the premade accompaniment tracks used by the artists [2]. Unlike the traditional song-writing and music production process where these two parts go hand in hand, communication and revision between beat makers and the rappers or the singers are not necessary. This new approach of music making is expanded out of hip-hop and adapted by the general music industry. American singer-songwriter Taylor Swift reveals her songwriting process of using pre-produced tracks in *I Wish You Would (Voice Memo)*[3]. Selling backing tracks on platforms like BeatStars[1] is an emerging business model which gains some significant success recently [4]. Because the mixing is merely between the main vocal and the accompaniment tracks, an automatic mixing system that focuses on this specific use case can either serve as a mixing assistant or finish the entire mix.

[1]BeatStars, http://www.beatstars.com/

B. Long-lasting Karaoke

Invented back in 1960's, karaoke has become a global phenomenon and brought joy to people for decades. With the rise of smartphones, mobile karaoke apps are soon getting popular, and users can share their singing on their social networks [5]. An automatic mixing system can improve the audio quality for users who have little knowledge of music mixing.

C. Mobile App Survey

Most karaoke mobile apps have some basic vocal mixing features, while a few of the apps provide the automatic mixing functionality. Besides the mobile karaoke apps, some music creation apps that target amateur users also provide similar features. The incomplete research investigates six popular karaoke apps (ChangBa[2], Smule[3], WeSing[4], PokeKara[5], StarMaker[6], and YoKee[7]) and two music creation apps (ChangYa[8] and Rap Fame[9]). Mixing setting templates and simplified parameters controls are the common designing choices. Balance control, reverberation, and time alignment are equipped in all apps. Graphic equalizers, noise gates, compressors, and special audio effects are also commonly seen. Pitch correction as a feature is only available for a portion of the song catalogue, and the vocal track notes are shifted to be centered around pitches of the pre-defined score. ChangBa[2] provides the automatic mixing service in terms of balance, equalization, and reverberation. From the author's personal experience, apps like Smule[3] may include some audio effects and automatic mixing function that are not visible to the users.

III. Related work / Context

A. Overview

Early automatic mixing methods are mostly rule-based expect systems for one type of audio effects. They incorporate some common practices by human mixing engineers into the

[2]ChangBa, https://changba.com/
[3]Smule, http://www.smule.com/
[4]WeSing, http://kg.qq.com/
[5]PokeKara, http://www.pokekara.com//
[6]StarMaker, http://www.starmakerstudios.com/
[7]YoKee, http://www.yokee.tv/
[8]ChangYa, http://www.i52hz.com/
[9]Rap Fame, https://rapfame.app/

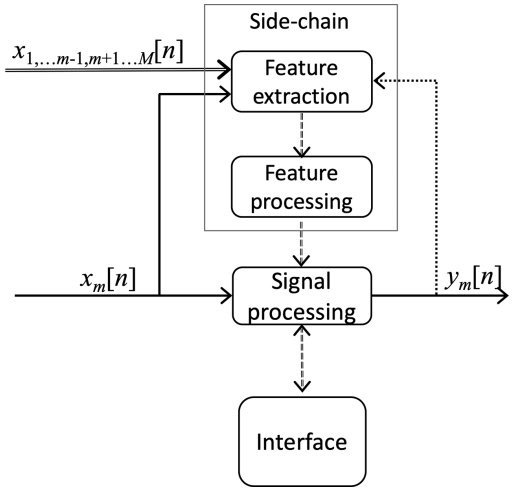

Fig. 1. A flow diagram of a typical feedback adaptive audio effect from [6]

algorithms that can perform similarly. Reiss [7] proposed the concept of adaptive audio effects. The parameters of an audio effects are modified by analysis of input audio through the feature extraction and mapping. This is the common approach for rule-based intelligent mixing. Figure 1. shows the diagram of a feedback-adaptive system which analyzes of the output audio to change the control parameters.

More about rule-based systems can be found in surveys where the previous research is grouped by audio effect types. A short historical overview by De Man et al. [8] presents the research timeline of automatic mixing from 2007 to 2017. Moffat and Sandler summarize the common methodologies to approach intelligent mixing in [1]. The book by De Man et al. [6] examines the challenges and the current solutions of intelligent music production comprehensively. The recent book chapter on automatic mixing by Moffat [9] includes the newest development in the data-driven approach.

B. Automatic Mixing in Pro Audio

Several audio products adapt the automatic mixing feature including the multi-effect plugin Neutron[10] and Balancer[11], and an online DAW Faders[12]. Their performances are not yet comparable with human mixes. However, in the task of automatic mastering, Landr[13] SoundCloud (powered by Dolby)[14], Ozone[15], BandLab[16], and eMastered[17] are generally considered successful for achieving the commercially usable quality.

[10]Neutron, https://www.izotope.com/en/products/neutron/features/mix-assistant.html
[11]Balancer, https://www.sonible.com/balancer/
[12]Faders, https://faders.io/
[13]Landr, https://www.landr.com/
[14]SoundCloud (powered by Dolby), https://community.soundcloud.com/mastering-on-soundcloud
[15]Ozone, https://www.izotope.com/en/products/ozone/features/master-assistant.html
[16]BandLab, https://www.bandlab.com/mastering
[17]eMastered, https://emastered.com/

C. Data-driven Approaches

The major limitation of the data-driven automatic mixing is the data collection. The common solution is to adapt the mix parameter reverse engineering approach [10] [11] as applied in [12] for drum level extraction. The parameter extraction requires the raw multi-track recordings and the corresponding professional mix-downs. This method is limited in the progress of reverse engineering research. One alternative is to let users create training data [13] [14].

It is possible to skip the process of extracting the mixing parameters by end-to-end data transformation. Martínez Ramírez and Reiss [15] present a proof-of-concept research using a deep autoencoder. Martínez Ramírez et al. [16] demonstrate a full end-to-end intelligent mixing system for drum tracks with Wave-U-Net[17]. This system learns the full mixing process of drum mixing and the result is indistinguishable from the professional engineer-generated mix. Besides the limitation of training data, another drawback of this method is that the end-to-end framework prevents further parameter control by humans. Steinmetz et al. [18] incorporate differentiable digital signal processing [19], and their system can produce human-readable mixing parameters. However, the implementation is very challenging, and the listening test result is not ideal. Martínez Ramírez et al. [20] propose a method to extract the mixing parameters for arbitrary black-box audio effects. This method does not need dedicated mix reverse engineering for specific audio effects nor differentiable digital signal processing implementation. The result is promising on two single-track tasks: non-speech sound removal and music mastering.

On the one hand, despite the significant progress of deep learning in intelligent mixing, the lack of data continues to be the barrier for creating a general-purposed automatic mixing system. One the other hand, the data-driven approach has shown convincing outcomes in specific tasks like drum mixing and single-track processing.

IV. Proposed method

Due to the scope of this project and the creative nature of music mixing, special audio effects (chorus, distortion, etc.) will not be considered. Time alignment, time correction and noise reduction will be temporarily ignored unless they are later proved to be necessities. Because the complexity of pitch correction, it will not be the focus of the research, but it will be implemented in the baseline system. The priorities are level balance, compression, equalization, and reverberation. For each, a rule-based baseline system and a data-driven system will be implemented.

A. Rule-based Baseline System

1) Level Balance: The assumption of level balance is that vocal-background ratio is determined by music genres. The vocal loudness analysis by genre from the work of the last semester will be used in combination with the genre classification. For genres that are not included, the vocal-background ratio will be set to the average value of the analysis.

2) Equalization: Automatic equalization can be achieved by frequency masking reduction [21]. The primary and the secondary frequency bands of each source signal are identified. And for each band, the masker source and the maskee source are determined. The bell filter cuts are applied to the masker in each frequency band. The parameter settings will be tuned for this specific mixing task.

3) Compression: An iterative process will optimize the compression threshold parameter to modify the dynamic range into a pre-defined target dynamic range within some tolerance. An alternative approach will be extract the dynamic profile of of the signal and compute the compression threshold to achieve the desired dynamic range as illustrated in figure 1. All the other parameters of the compressor are fixed. The dynamic range measurement will be the EBU dynamic range [22].

4) Reverberation: In [23], a set of mixing rules that maps tempo and crest factor into parameters including gain, pre-delay, diffusion, tail decay, high frequency damping, and high frequency cut. The rules will be simplified and modified to serve the vocal mixing task. An intelligent equalizer can be applied after reverberation in the signal chain for mask reduction.

5) Pitch Correction: A pitch corrector plugin will be used, and the setting parameters are either fixed or determined by key detection and chord detection.

B. Proposed Data-driven System

Because there is no available dataset for raw singing vocal and mixed vocal tracks, the previous data-driven approaches that directed extract the mixing parameters are not applicable. Two proposed solutions are 1)estimating the intermediate audio features from the processed signals for adaptive audio effects and 2) generating the raw recordings form the processed signal.

The training data is clean professional or semi-professional produced vocal tracks and backing tracks. More lower quality data can be collected from source separation of commercial recordings. The input to the neural network will be the audio signals of both the vocal and backing tracks in either time signal, magnitude spectrogram, or mel-spectrogram format. The ground truth will be the mixing parameters and the targeted intermediate audio features. The input signal will be altered (normalization, de-reverberation, expansion, etc.) in pre-processing to ensure the the ground truth cannot be calculated from the input. In the cases of level balance, compression, and reverberation, the vocal signal can be omitted in training. The input audio length can be set to 3 seconds. During inference, the model output may be smoothed by a moving-average filter.

1) Level Balance: The ground truth is the relative loudness between the vocal track and the backing track. During the inference time, the vocal gain will be calculated based on the current relative loudness and the model output. The loudness measurement will be either the short-term loudness or the programme loudness [24]. Silence will be considered.

2) Equalization: Given the processed vocal signal available, the "raw" vocal signal can be generated by randomly manipulating the processed signal with equalizers. The model is then trained to find the proper equalizer parameters that recover the original signal. Because the manipulation settings are known, there is no need for the reverse engineering framework to extract the ground truth data.

3) Compression: The ground truth is the dynamic range of the vocal signal. The deep learning model outputs a desired dynamic range value based on the input signals. The optimization process will be the same as it is in the rule-based system.

4) Reverberation: Reverberation parameter extraction is a challenging task. Some commercial products are able to match the reverberation characteristics of voice recordings with algorithmic reverberation. However, it is yet unknown how accessible the internal algorithmic reverberation processing and the extracted parameters are, although the estimated impulse responses are guaranteed. An alternative is to use genetic algorithms to approximate the impulse responses with another algorithmic reverberation processor [25]. A formal objective evaluation of the reverberation parameter extraction will be conducted. Because the use of the parameter extraction is to support the ground truth data creation, errors within a limit are acceptable.

5) Pitch Correction: An existing framework Deep Auto-tuner[26] can be incorporated and improved.

V. Proposed Evaluation

At least twenty participants will be recruited as volunteers for the subjective listening test. The participants should have some critical listening experiences. The participants will be mixed with males and females. The ages will be recorded. The test will be no longer than 30 minutes to avoid listening fatigue [27].

Participants will complete the online listening experiment constructed with either Qualtrics[18] or the Web Audio

[18]Qualtrics, https://www.qualtrics.com/

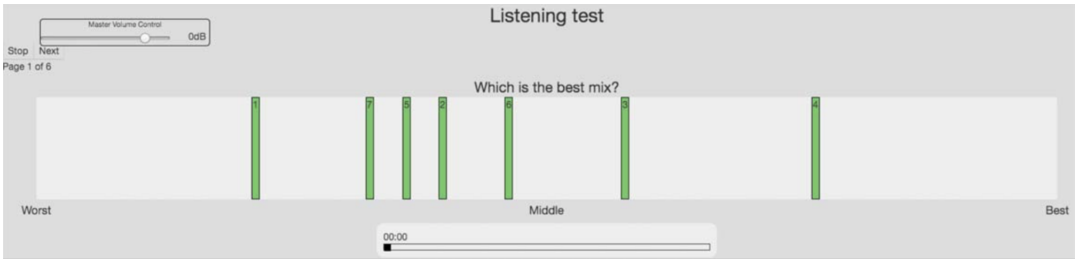

Fig. 2. Listening experiment interface from [16]

Evaluation Tool [28]. Participants are required to conduct the experiment in critical listening environment or with a good pair of quality headphones in low-noise environment. Participants will adjust the volume to a comfortable level at the beginning of the experiment and stay in this level during the experiment. Fig. 2 is an example of the experiment interface. The participants will rate the audio samples in a continuous scale. Participants will be asked to rate their preference on five different mixes of the same source materials. One of the listening sample will be mixed by a professional engineer. One will be the unprocessed summing track of the raw tracks. One will be the output of a commercial automatic mixing system. One will be the output of the rule-based baseline system. One will be the output of the proposed data-driven automatic mixing system. The raw recording tracks are from either the DAMP-VSEP dataset [29] or the MIR-1K dataset [30].

All listening samples will be normalized to 28 dBFS from ITU-R BS.177-2 [24], or other appropriated listening loudness levels. The order of the listening samples will be randomized to avoid bias. All audio samples used for evaluation will be placed online after the experiment if permitted.

Tests will be conducted to prove the existence of statistically significant difference by comparing the p-value and the α-value.

VI. Novelty of proposed work

Firstly, the proposed work identities a new problem that has real use cases but has not been addressed by the research community. Based on the recent progress of deep learning in intelligent mixing on in the single-track context, the limited increase of complexity of the proposed work has some potential to success.

Secondly, due to the lack of datasets, the proposed work provides solutions to generate training data from the human-mixed audio for each audio effect type without the need of raw recordings. For level balance and compression, this work proposes to extract the intermediate parameters as the ground truth, bridging the knowledge-based feedback adaptive audio effects and the machine learning approach. It further proposes the possibility to extend the dataset with source separation.

VII. Required resources

A. Datasets

For model training and testing, there are several datasets providing processed singing vocal and backing tracks. The Rock Band dataset [31] has more than 60 full-length professional-mixed stereo commercial recordings in the multi-track format. The instrumental tracks are balanced in level and can be summed to form the backing tracks. The MUSDB18HQ dataset [32] comprises 150 full-length stereo tracks from Mixing Secrets[19] and MedleyDB [33], and processed vocal and instrumental stems are provided. The iKala dataset [34] includes the singing vocal and backing tracks in 252 30-second mono audio samples. The ccMixter dataset[35] consists of 50 full-length stereo singing vocal and backing tracks mixed by amateur engineers. Besides the existing datasets, source separation can generate almost infinite amount of data from commercial releases with tolerable interference and artifacts.

Raw singing vocals and corresponding backing tracks are needed for the final evaluation. The DAMP-VSEP dataset [29] contains cellphone recorded amateur singing and backing tracks of 11494 songs. The MIR-1K dataset [30] consists of 133 min unprocessed mono singing vocal and backing tracks.

B. Tools

This research will be conducted in Python environment.

1) Third-party Libraries: The Essentia library [36] and the librosa libaray [37] will assist the implementation of the rule-based system. The Pytorch library [38] or the Tensorflow [39] library will be the framework to implement the deep learning model. Some existing models like DeepAFx [20] for the similar tasks may be borrowed. The source separation package Spleeter[40] may be used to generate training data. Commercial products like Chameleon[20] and Dialogue Match[21] may be used to extract reverberation parameters. A python audio plugin host (RenderMan [41], DawDreamer [42], or Pedalboard [43]) will be part of the system.

2) Audio Plugins: IEM Plug-in Suite[22] is a series of audio plugins with source code available, including equalizers, gates, compressors, and reverbs. Dragonfly reverb[23] is an open-

[19]Mixing Secrets, https://www.cambridge-mt.com/ms/mtk/
[20]Chameleon, https://www.accentize.com/chameleon/
[21]Dialogue Match, https://www.izotope.com/en/products/dialogue-match/features/reverb.html
[22]IEM Plug-in Suite, https://plugins.iem.at/
[23]Dragonfly Reverb, https://github.com/michaelwillis/dragonfly-reverb

source reverberation plugin. MAutoPitch[24] is a free pitch correction plugin.

VIII. Deliverables

The deliverable of this project will be an offline command-line python program. Given the raw singing vocal track and the backing track, the program can automatically mix the two tracks and output the mixed audio. The mixing quality is aimed to surpass the baseline system and approach the human mixing.

IX. Timeline

February	1) signal processing chain in Python 2) rule-based level balance 3) reverb parameter extraction search
March	1) rule-based compression 2) rule-based equalization 3) rule-based reverberation 4) data collection
April	1) finalized rule-based system 2) data collection 3) data-driven level balance
Summer	1) data collection
September	1) data-driven compression 2) data-driven equalization
October	1) data-driven reverberation 2) experiment preparation
November	1) experiment
December	1) paper writing

References

[1] D. Moffat and M. B. Sandler, "Approaches in intelligent music production," in *Arts*, Multidisciplinary Digital Publishing Institute, vol. 8, 2019.

[2] J. G. Schloss, *Making beats: The art of sample-based hip-hop*. Wesleyan University Press, 2014.

[3] T. Swift, *I wish you would (voice memo)*, 2014. [Online]. Available: https://www.youtube.com/watch?v=O8DBKZSQ2Cw.

[4] M. Stassen, *Having paid out $150m to creators, BeatStars launches Sony Music Publishing-backed publishing service*, 2021. [Online]. Available: https://www.musicbusinessworldwide.com/having-paid-out-150m-to-creators-beatstars-launches-sony-music-publishing-backed-publishing-service/.

[24] MAutoPitch, https://www.meldaproduction.com/MAutoPitch

[5] X. Zhou and F. Tarocco, *Karaoke: The global phenomenon*. Reaktion Books, 2013.

[6] B. De Man, R. Stables, and J. D. Reiss, *Intelligent Music Production*. Audio Engineering Society Presents. Taylor Francis, 2019.

[7] J. D. Reiss, "Intelligent systems for mixing multichannel audio," in *2011 17th International Conference on Digital Signal Processing (DSP)*, IEEE, 2011.

[8] B. De Man, J. Reiss, and R. Stables, "Ten years of automatic mixing," in *3rd Workshop on Intelligent Music Production*, vol. 15, 2017.

[9] D. Moffat, "Ai music mixing systems," in *Handbook of Artificial Intelligence for Music*, Springer, 2021, pp. 345–375.

[10] D. Barchiesi and J. Reiss, "Reverse engineering of a mix," *Journal of the Audio Engineering Society*, vol. 58, no. 7/8, 2010.

[11] J. T. Colonel and J. Reiss, "Reverse engineering of a recording mix with differentiable digital signal processing," *The Journal of the Acoustical Society of America*, vol. 150, no. 1, 2021.

[12] D. Moffat and M. Sandler, "Machine learning multitrack gain mixing of drums," in *Audio Engineering Society Convention 147*, 2019.

[13] E. T. Chourdakis and J. D. Reiss, "Automatic control of a digital reverberation effect using hybrid models," in *Dereverberation and Reverberation of Audio, Music, and Speech Conference*, Audio Engineering Society, 2016.

[14] E. T. Chourdakis and J. D. Reiss, "A machine-learning approach to application of intelligent artificial reverberation," *Journal of the Audio Engineering Society*, vol. 65, no. 1/2, 2017.

[15] M. A. Martınez Ramırez and J. D. Reiss, "Deep learning and intelligent audio mixing," in *3rd Workshop on Intelligent Music Production*, vol. 15, 2017.

[16] M. Martinez Ramirez, D. Stoller, and D. Moffat, "A deep learning approach to intelligent drum mixing with the wave-u-net," Audio Engineering Society, 2021.

[17] D. Stoller, S. Ewert, and S. Dixon, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," in *Proceedings of ISMIR*, 2018.

[18] C. J. Steinmetz, J. Pons, S. Pascual, and J. Serrà, "Automatic multitrack mixing with a differentiable mixing console of neural audio effects," in *ICASSP*, IEEE, 2021.

[19] J. Engel, C. Gu, A. Roberts, *et al.*, "DDSP: Differentiable digital signal processing," in *International Conference on Learning Representations*, 2019.

[20] M. A. Martínez Ramírez, O. Wang, P. Smaragdis, and N. J. Bryan, "Differentiable signal processing with black-box audio effects," in *ICASSP*, IEEE, 2021.

[21] S. Hafezi and J. D. Reiss, "Autonomous multitrack equalization based on masking reduction," *Journal of the Audio Engineering Society*, vol. 63, no. 5, 2015.

[22] E. T. 3342, "Loudness range: A measure to supplement ebu r 128 loudness normalisation," 2016.

[23] D. Moffat and M. Sandler, "An automated approach to the application of reverberation," in *Audio Engineering Society Convention*, 2019.

[24] EBU–Recommendation, "Loudness normalisation and permitted maximum level of audio signals," 2011.

[25] S. Heise, M. Hlatky, and J. Loviscach, "Automatic adjustment of off-the-shelf reverberation effects," in *Audio Engineering Society Convention*, 2009.

[26] S. Wager, G. Tzanetakis, C.-i. Wang, and M. Kim, "Deep autotuner: A pitch correcting network for singing performances," in *ICASSP*, IEEE, 2020.

[27] R. Schatz, S. Egger, and K. Masuch, "The impact of test duration on user fatigue and reliability of subjective quality ratings," *Journal of the Audio Engineering Society*, vol. 60, no. 1/2, 2012.

[28] N. Jillings, B. D. Man, D. Moffat, J. D. Reiss, *et al.*, "Web audio evaluation tool: A browser-based listening test environment," 2015.

[29] I. Smule, *DAMP-VSEP: Smule digital archive of mobile performances - vocal separation*, Oct. 2019. [Online]. Available: https://doi.org/10.5281/zenodo.3553059.

[30] C.-L. Hsu and J.-S. R. Jang, "On the improvement of singing voice separation for monaural recordings using the MIR-1K dataset," *IEEE transactions on audio, speech, and language processing*, vol. 18, no. 2, 2009.

[31] *"Rock band" (video game)*, Harmonix, MTV Games, Electronic Arts, 2008.

[32] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, *MUSDB18-HQ - an uncompressed version of musdb18*, Dec. 2019. [Online]. Available: https://doi.org/10.5281/zenodo.3338373.

[33] R. M. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam, and J. P. Bello, "MedleyDB: A multitrack dataset for annotation-intensive mir research.," in *ISMIR*, vol. 14, 2014.

[34] T.-S. Chan, T.-C. Yeh, Z.-C. Fan, *et al.*, "Vocal activity informed singing voice separation with the ikala dataset," in *ICASSP*, IEEE, 2015.

[35] A. Liutkus, D. Fitzgerald, Z. Rafii, B. Pardo, and L. Daudet, "Kernel additive models for source separation," *IEEE Transactions on Signal Processing*, vol. 62, no. 16, 2014.

[36] D. Bogdanov, N. Wack, E. Gómez Gutiérrez, *et al.*, "Essentia: An audio analysis library for music information retrieval," in *ISMIR*, 2013.

[37] B. McFee, C. Raffel, D. Liang, *et al.*, "Librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, vol. 8, 2015.

[38] A. Paszke, S. Gross, F. Massa, *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[39] M. Abadi, P. Barham, J. Chen, *et al.*, "Tensorflow: A system for large-scale machine learning," in *12th USENIX symposium on operating systems design and implementation (OSDI 16)*, 2016.

[40] R. Hennequin, A. Khlif, F. Voituret, and M. Moussallam, "Spleeter: A fast and efficient music source separation tool with pre-trained models," *Journal of Open Source Software*, vol. 5, no. 50, 2020. [Online]. Available: https://doi.org/10.21105/joss.02154.

[41] L. Fedden, *RenderMan*, 2017. [Online]. Available: https://github.com/fedden/RenderMan.

[42] D. Braun, *DawDreamer: Bridging the Gap Between Digital Audio Workstations and Python Interfaces*. [Online]. Available: https : / / github . com / DBraun / DawDreamer.

[43] Spotify, *Pedalboard*. [Online]. Available: https://github.com/spotify/pedalboard.